Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 7, No. 1, June 2019, pp.395-407
Available online at: http://pen.ius.edu.ba

Real time data analysis and visualization for the breast cancer disease

Sanyour Rawan[1], Abdullah Manal[2]
[1,2]Department of Information System, King Abdul Aziz University

Article Info

Received Nov 11, 2018

Keyword:

Data visualization
Interactive data visualization
Shiny app
Breast cancer prediction
Classification

ABSTRACT

Today, the amount of data that are digitally collected in the healthcare sector is tremendous and expanding rapidly, these data are inherently geospatial and temporal ranging from individual families to whole states and from minutes to decades. Therefore, they need sophisticated data management and analysis to be transformed into valuable knowledge. Healthcare professionals are faced with several challenges regarding extracting knowledge from this massive amount of data in order to support the decision-making process. To gain advantage of health care big data, big data analytics need to be exploited to utilize and understand patterns associations within these data thus make the right decision. In this research, an interactive data analysis and visualization tool is proposed to visually compare the performance of three machine learning algorithms on Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The proposed model consists of two phases: input phase and analysis/visualization phase. It aims to allow the user to interactively compare the performance of three different ML algorithms (KNN, SVM and NB) in terms of accuracy, sensitivity and error rate in a user-friendly way. Here, SVM classifier has proven its efficiency and it is concluded as the best classifier with the highest accuracy as compared to the other two classifiers.

Corresponding Author:

First Author,
Department of Information System,
King Abdul Aziz University,
Jeddah, Saudi Arabia.
Rsanyour0001@stu.kau.edu.sa

1- Introduction

Problems solving and decision making in the healthcare sector is extremely depending on accessing and managing knowledge. Today, with tremendous amounts of data generated every day, it has become essential for healthcare organizations to efficiently manage and process their knowledge in order to provide their best care services, accomplish operational excellence and enhance the decision-making process. Effective well-organized knowledge management approaches can assist these organizations to achieve such goals. In the healthcare sector, knowledge management systems can be defined as tools that can provide critical information regarding characteristics or circumstance of clinical situations. Despite the fact that their results required to be interpreted by humans to be applicable to a specific patient, these systems provide a wide range of strategies and resources to effectively create, present, interpret and share of medical knowledge to facilitate the flow of information and result in better and more-informed decisions in clinical practices (Lobach et al, 2012).

Data visualization is one of the fundamental knowledge management aspects that can provide a precise view of data and allow researchers to gain deep insight in order to improve research outcomes and make better decisions (Ko and Chang, 2018). Data visualization can effectively present substantial results summaries, identify hidden patterns in data and reveal data quality problems resulted by using data created for other

purposes (Plaisant et al., 2014). In addition, it can significantly help to identify patients who match clinical trials selection criteria, monitor and control the spread of infectious diseases. Moreover, by using visual analytics, patients can monitor their own conditions, review treatment plans, gain social support and communicate with their care team (Shneiderman et al., 2013).

With the unfortunate increasing trend of breast cancer cases and with the urgent need of handling massive amount of data in medical and clinical researches, the utilization of knowledge management approaches and data mining techniques in medical sector is growing rapidly due to their efficiency in classifications and predication processes. Machine learning approaches have a significant impact in improving clinical studies, reducing medicine cost, enhancing patient outcomes and helping medical professionals in decision making processes (Salama et al., 2012).

Breast cancer is the most common malignancy among women with nearly 1 in every 3 cancer diagnose in united states, thus, it becomes the second leading cause of cancer death among women, according to (Web-1). Worldwide, and according to World Cancer Research Fund, Breast cancer is the most commonly occurring type of cancer in woman and the second common cancer overall (Web-2).

Breast cancer occurs when the breast tissue cells start to grow abnormally performing a mass commonly referred to as a Tumor. The existence of the tumor does not indicate that it is cancer, instead, tumors can be benign (not cancerous), or malignant (cancerous).

Currently, several tests, including mammogram, FNA (Fine Needle Aspiration) and biopsy can be performed to diagnose breast cancer and discover whether the tumor is benign or malignant. A breast FNA is a medical test procedure by which cells are extracted from the tumor or lump by a needle similar to that used in blood sample test. This sample will be sent to the laboratory to be examined by a specialist doctor called cytologist to determine the nature and characteristics of that lump (Mendoza et al., 2011).

In this research, a dataset constitutes of some tumor characteristics detected by the FNA test along with Machine Learning Algorithms and Shiny app package are used to build an interactive tool that can significantly provide a computational interpretation and interactive data visualization that can be changed according to parameters' changes. By using such tool, scientists can analyze the data visually, discover tumor characteristics and the correlation between them and find whether if there is a "causation" relation between specific features. It can be used as a real-time breast cancer detection tool.

The rest of this paper is organized as follows. Section 2 is background and section 3 is the related work. Section 4 illustrates the interactive data analysis/visualization model, compares and discusses experiments results obtained. Finally, section 5 concludes the research.

2- Background

In this section, a brief description of interactive data visualization and R data visualization packages is provided.

2.1- Interactive Data Visualization

Data visualization is an essential part of data analytics. It refers to using tables and graphs and other kinds of figures to represent data in an efficient way by which people can understand data in a graphical manner. Visualization is the best way to explore and discover hidden patterns of data. By visualizing data, the required information can be disposed on a single screen, thus users can easily understand data and accordingly, the right decisions can be taken easily and precisely. Exploring data using different graphs enables humans' eyes to easily detect meaningful information, understand large datasets, identify meaningful patterns within data, recognize outliers and form hypothesis efficiently

Unlike traditional data that are well organized and structured, big data such as streaming data, images, documents, videos and music are complex, unstructured and unorganized thus required advanced visualization techniques rather than the traditional static ones. Since that, traditional techniques for presenting the tremendous growing amount of data have some limitations, dynamic data visualization became a significant alternative to effectively communicate and display datasets in an accessible manner not only for experts but also for non-experts, who can perform complex statistical analysis using a "point- and-click interface", as well (Ellis and Merdian, 2015)(Agrawal et al., 2015)(Cho et al., 2014).

Wang et al. (2015), stated four primary steps for interactive data visualization: 1) Selecting: interactively choosing a subset of the whole dataset according to the user's interests. 2) Linking: relating information among different multiple views. 3) Filtering: focusing on the data of interest by adjusting the amount of data

to display in order to decrease information quantity. 4) Rearranging: rearranging and re-mapping the layout of displayed information to bring different insights.
According to Ellis and Merdian (2015), dynamic data visualization has a significant advantage in teaching complex statistical concepts to improve students' understanding. In addition, in some applied sciences such as clinical psychology, dynamic data visualization can successfully attain some scientists' research requirements. Frequently, in large datasets, there are many variables that cannot be properly presented by standard linear representation. Moreover, dynamic data visualizations allow non-experts to repeat sophisticated analyses and they still get same results.

2.2 – Why Shiny App Framework?

R is a significant tool for Big data analytics research. Although that R does not provide interactive services, and because that data visualization is an essential part of Big data analytics, R users and researchers have been developed various packages that provide effective interactive functionalities. Web interface packages that enable the interactivity over the web, and graphic packages which provide graphical visualization functions in R are examples of such packages. some of these R interactive and visualization function packages are RgoogleMap, GoogleVis and Shiny.
RgoogleMap (Loecher and Ropkins, 2015) is a spatial Big data analysis Package by which users can query Google server for static maps. These maps are used as a background images to present plots within R. using this package requires accurate coordinates scaling.
GoogleVis (Gesmann and Castillo, 2011) provides an intermediate interface between R and Google Chart Tools GCT API. By using the GCT, and based on R data frames, users can create web pages with interactive charts capabilities. These charts can be displayed via R HTTP local server or directly within users' sites without uploading their data to Google.
Shiny (Web-3) is one of the effective interactive visualization packages provided in R that allows statisticians and data analysts to share their work over the web in an interactive way. Shiny apps are accessible by anyone can interact with them regardless their statistics and analysis background. They allow users to interact with data in real time and generate customized graphs based on their needs and focus on what they want to see from their own perspective.
Shiny web applications have several advantages: 1) they don't require a knowledge in web development. 2) They are user friendly. 3) Can be accessed from any device. 4) They are highly customizable and allowing for user modifications. (Scrivner et al., 2017)

3- Related work

Several prior studies have proposed the application of machine learning algorithms in breast cancer classification using the (WDBC) dataset, these studies showed significant results. Asri et al.(2016), compared the performance of a group of ML algorithms: Support Vector Machine, Decision Tree (C4.5), Naive Bayes and k Nearest Neighbors on Wisconsin Breast Cancer dataset to find the most accurate algorithm with the lowest error rate. Their experiment showed that SVM achieved the best accuracy with 97.13 and error rate .02% ,while the accuracy in KNN and NB varied between 95.12% and 95.28%. To gain the best accuracy in breast cancer patients' identification and thus increase breast cancer survivability, on 2014, Chaurasia and Pal have proposed a diagnosis system to evaluate the performance of RepTree, RBF Network and Simple Logistic algorithms, their study concluded that the Simple Logistic algorithm has approved to be the best classifier among the three classifiers with 74.47% accuracy (Chaurasia and Pal, 2014). Classifiers such as Naïve Bayes, SVM-RBF kernel, Radial basis neural networks, Decision trees J48 and CART have been used by Aruna et al. to discover the best classifier in detecting breast cancer by comparing their accuracy, sensitivity, precision and specify. On WBC dataset and by using WEKA, their results indicated that SVM RBF registered the highest accuracy with 99% (Aruna et al., 2011). Salama et al. (2012) have proposed a comparison among various classifiers, including decision tree (J48), Multi-Layer Perception (MLP), Naive Bayes, Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest neighbor (IBK). Based on 10-fold cross validation technique to increase the accuracy, their experiments applied on three different datasets: Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) to compare accuracy and confusion matrices. To gain optimal results, they introduced a combination between different classifiers at the classification process to find the most appropriate multi classifier approach for each proposed dataset. On the WDBC dataset, results showed that using the fusion of Sequential Minimal Optimization and Multi-Layer Perception or using only Sequential Minimal Optimization are superior to other

classifiers. On his proposed paper, Rodrigues used two machine learning algorithms: Bayesian Networks and J48 to create classifiers by which benign and malignant tumors could be discriminated accurately. In order to optimize these classifiers, a considerable work was dedicated to preprocess the data. The best accuracy was achieved by Bayesian Networks algorithm with 97.80% (Rodrigues, 2016)

This paper introduces another study on this topic while using an interactive data analysis and visualization tool that allows the user to interactively visualize the dataset features and inspect the correlation between attributes, compare the performance of three different classifiers: KNN, SVM and NB. In addition, the proposed tool allows the user to visually adjusting the classifiers' parameters such the value of K in KNN and Cost, Gamma in SVM to gain the best accuracy.

4 – Interactive Data Analysis/Visualization Model

As shown in figure1, this Interactive data Analysis/visualization Model has two primary phases: input phase and analysis/visualization phase. The first phase is the input phase, which is responsible for allowing the user to upload his own data through the application user interface while the second phase is responsible for analyzing and visualizing the uploaded data. The sequence of the model operation is depicted on figure 2 where the user can upload the dataset after the application run for the first time. Then, the data will be automatically visualized. If one of the parameters' values changed by the user, the changes will be reflected directly on the screen accordingly. In the following subsections, the two phases of the proposed model will be described in detail.

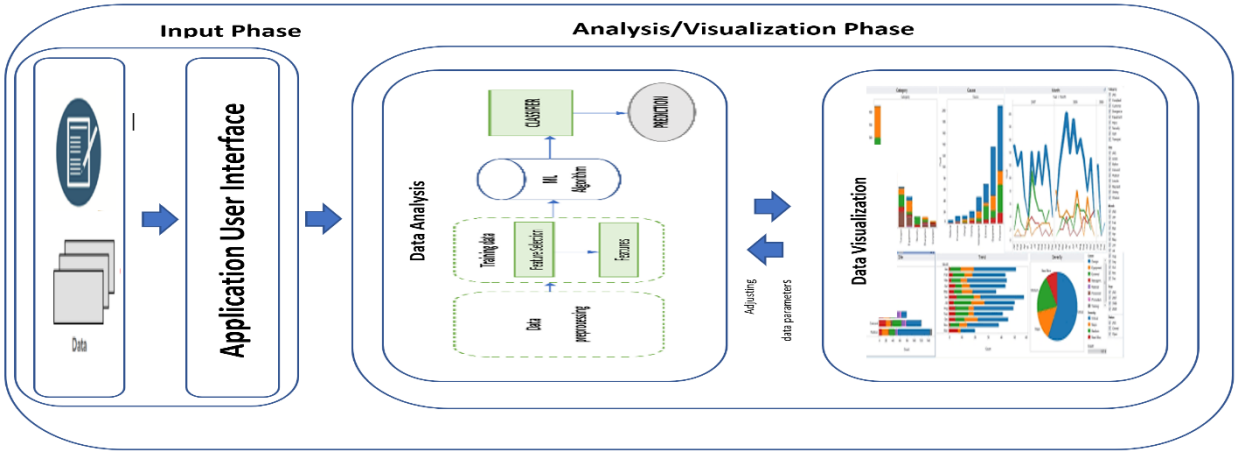

Figure1 Data Analysis/Visualization Application Model.

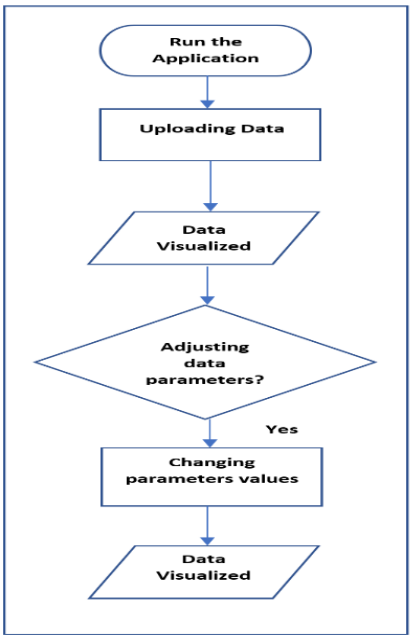

Figure2 Analysis/VisualizationFlowchart

4.1 – Input Phase

As mentioned earlier, this model is designed to provide an interactive data analysis/visualization tool by which users can upload their data to visually discover its features and analyze data with a user-friendly design. After the application run for the first time, the user can use data bar as shown in figure3 to upload the dataset, discover his dataset features visually, adjust parameters settings and the results will be reflected accordingly on the screen based on selected features and parameters' values in the real time.

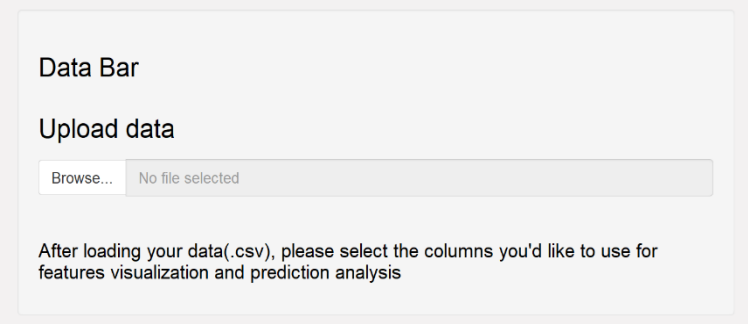

Figure3 Uploading the Dataset

The application as illustrated in figure4, consists of several tabs: "data visualization" tab which can be used to discover dataset's features in an interactive way, "KNN classifier" tab by which users can apply the K Nearest Neighbor machine learning algorithm to their data to examine the impact of changing K value on the KNN prediction accuracy, the third one is "SVM classifier" tab, this tab will allow the user to tune the fundamental Support Vector Machine algorithm parameters such as gamma and Cost to achieve the highest accuracy level and the fourth tab is "NB Classifier" which is enables the user to apply the Naive Bayes algorithm on the uploaded dataset.

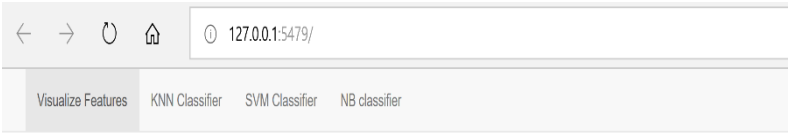

Figure4 Four Different Tabs in The Application

4.1.1- Dataset Description

The Breast cancer Dataset was created by Dr. William H. Wolberg, a physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA and it is publically available. It constitutes of 11 attributes (including patient ID) and 699 instances. Dr. Wolberg collected these samples from a group of patients with solid breast masses and performed an analysis of cytological features based on the tumor digital scan by using a graphical computer program called Xcyt. The program used a curve-fitting algorithm to compute ten different features from each cell in the sample set. Table1 describes these features (Rodrigues, 2016) (Web-1) (Web-4) .

Table 1 Breast Cancer Dataset Features

Feature	Datatype	Range
“Clump Thickness”	Numeric	1-10
“Uniformity of Cell Size”	Numeric	1-10
“Uniformity of Cell Shape”	Numeric	1-10
“Marginal Adhesion”	Numeric	1-10
“Single Epithelial Cell Size”	Numeric	1-10
“Bare Nuclei”	Numeric	1-10
“Bland Chromatin”	Numeric	1-10
“Normal Nucleoli”	Numeric	1-10
“Mitoses”	Numeric	1-10
“Class”	Factor	“Benign”, “Malignant”
NO. of missing values (NAs)	16	
Total NO. of instances	699	

On a scale of 1 to 10, each feature is given a value with 1 being “benign” and 10 being “malignant”. These nine features (mentioned in table1) are significantly varies between benign and malignant samples.
This dataset is considered as “noise-free” and contains 16 missing values in 16 different instances, all of them are in Bare Nuclei attribute. At the first stage, data preprocessing step will focus on how to deal with these missing values. To manage this issue, scientists commonly use one of two methods: the first one is replacing each missing value with the mean of that attribute values in training data. The second is to remove all the records that have missing values (Rodrigues, 2016). In this research, the second option will be adopted, so the new dataset will contain 683 instances.

4.2 – Analysis/Visualization Phase

This phase is divided into two fundamental steps: Data analysis step, and data visualization step. The following subsections demonstrate these steps in detail.

4.2.1 – Data Analysis

In “KNN classifier” tab, the KNN classification algorithm will be applied on the dataset and the confusion matrix will be displayed for the user. Based on his selection of K value (given range value 1-20) and the attributes to be included in the prediction process as shown in figure5, results will be changed automatically after each new parameter adjustment.

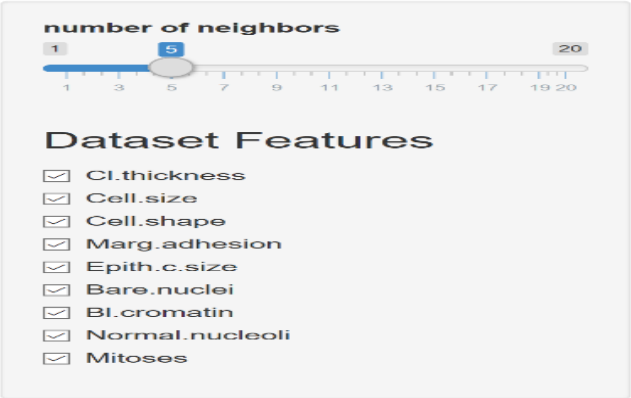

Figure5 NO. of Neighbors and Dataset Features Selection

Confusion matrices are the basic indicators that used to evaluate classification errors and thus measure the performance of machine learning classification. They illustrate the misclassifications occurring during the classification process, i.e., they compare the number of misclassified (erroneously predicted) items in each

class with the actual class in which items should be classified. As shown on table2, the confusion matrix generates four different values: TP, FP, FN and TN.

- TP: True Positive: the prediction is positive, and it is true.
- TN: True Negative: the prediction is negative, and it is true.
- FP: False Positive (Type 1 error): the prediction is positive, and it is false.
- FN: False Negative (Type 2 error): the prediction is negative, and it is false.

Table 2 Confusion Matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

The selection of parameters setting (k value and attributes included) is basically depending on the tolerance to Type1 and Type2 errors. In some domains Type1 error is critical while Type2 is more tolerated. In situations of medical diagnosis as in the proposed application, Type2 error is critical (the false negative), it means that the classifier will classify a malignant tumor as a benign thus, the correct diagnosis will be delayed (Beauxis-aussalet and Hardman, 2016) Web-5.
The second is SVM algorithm which is applied on the dataset in order to examine its classification performance. There are several learning parameters that can be tuned in constructing SVM for regression. Two essential tuning parameters by which the classification accuracy can be considerably increased. These two parameters are: Gamma and Regularization parameter (C). Varying their values, more accuracy can be achieved in a reasonable amount of time. C parameter indicates how much the SVM optimization should avoid misclassifying each training data point. By changing the parameter C, the trade-off between the complexity of decision rule (the model) and frequency of error can be controlled. Increasing and decreasing the value of C parameter will change the margin space around the separation line (hyperplane).
Gamma parameter defines the degree of influence that a single training data point can reach. Decreasing Gamma means that points located far away from the separation line are also considered in calculations while by increasing It, only points that are "close" to the separation line will be considered, therefore, it can be concluded that reducing Gamma will negatively affect the classification accuracy. At the first time the app ran, these parameters are tuned to be 10 for C and 1 for Gamma as illustrated in figure6.

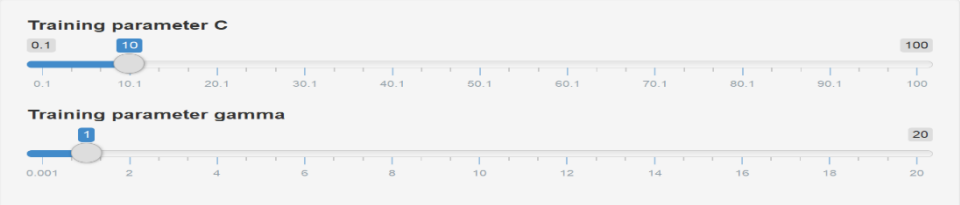

Figure6 Initial C and Gamma Values

The third is the Naïve Bayes algorithm which is applied on the dataset in order to examine its classification performance. This algorithm is based on the so-called Bayesian theorem, thus, it allows to predict a class y, given a set of features x1, …., xn by using probability. The class with the highest probability will be considered as the most occurrence class, this feature is called "**Maximum A Posteriori (MAP)**".

4.2.2 – Data Visualization

To validate the application functionality, five different experiments are performed in which machine learning algorithms are applied on breast cancer dataset to discover this interactive data analysis/visualization tool capabilities.

Experiment1. In this experiment, the interactive data visualization capabilities of the proposed application are discovered. At the first time that the application ran, at the "visualize features" tab, three different graphs are displayed as shown in figure7.

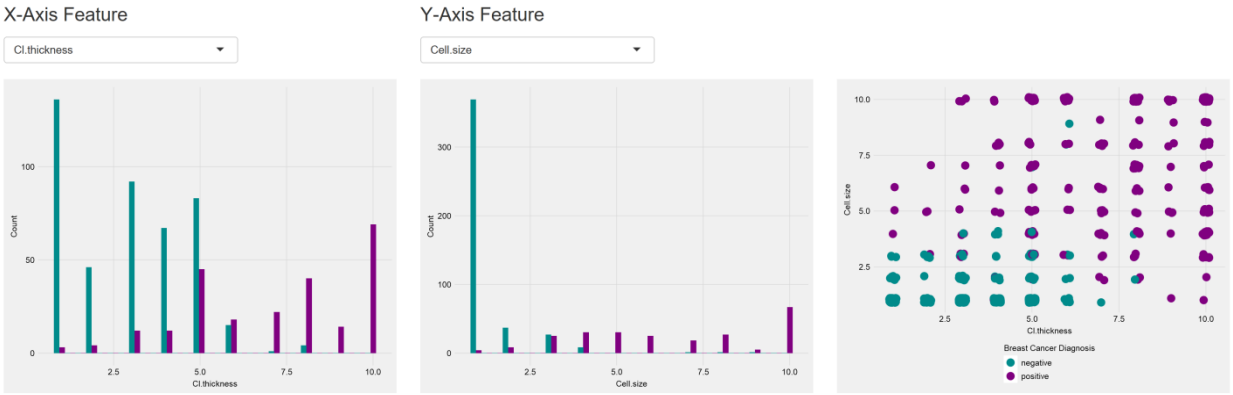

Figure7 Initial Displayed Graphs

As default values, Cell thickness and Cell size features are selected to represent X-Axis and Y Axis respectively. The first and second graphs illustrate how these two features differ significantly between benign and malignant samples. The third graph shows the relation between Cell thickness and Cell size. As the values of these two features increase, the possibility that the tumor well be malignant is increased as well. Next, the X-axis feature is changed to be Cell shape VS Cell size.

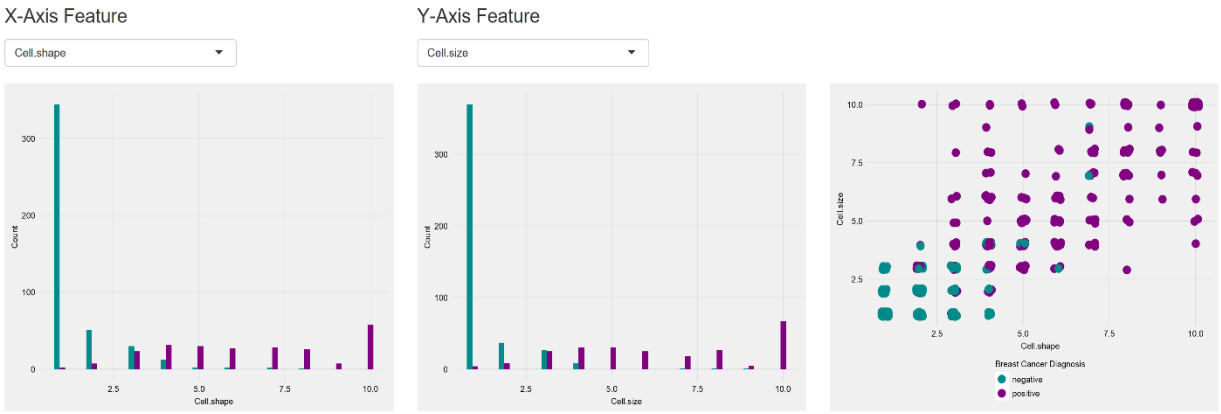

Figure8 Relation Between "Cell.shape" and "Cell.size"

As shown in figure8, the cancer cells tend to vary in terms of size and shape. The value 5 of cell size can be considered as a boundary line between being a benign or a malignant tumor. If cell size is less than 5, based on the Cell shape, the tumor could be either benign or malignant while above 5, the tumor is overwhelmingly diagnosed as malignant. However, these observations cannot be generalized since that each type of cancer has its own aspects and the behavior of the malignant cells can be measured by several other parameters such as the cell ability to find a suitable environment for its growth, the state of the tumor infiltrating immune cells and the interaction of the environmental factors of that tumor. Figure9 indicates that as the value of "Epithelial Cell Size" increased, the lump is tending to be malignant.

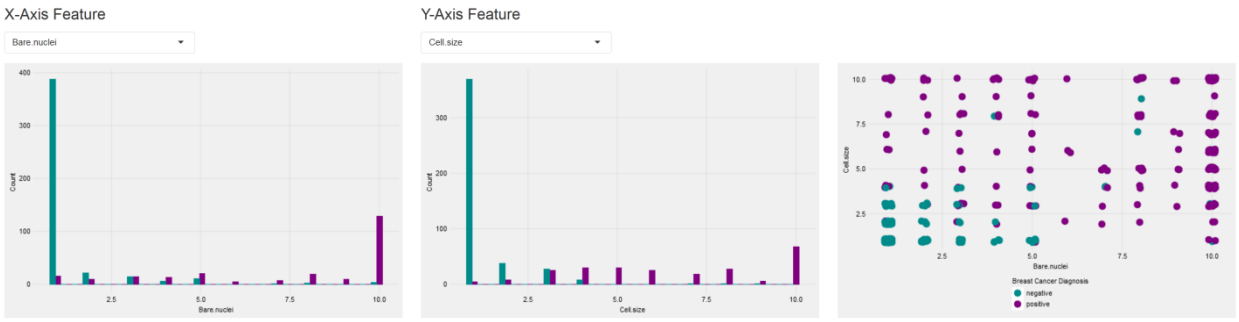

Figure9 Relation Between "Bare.nuclei" and "Cell.size"

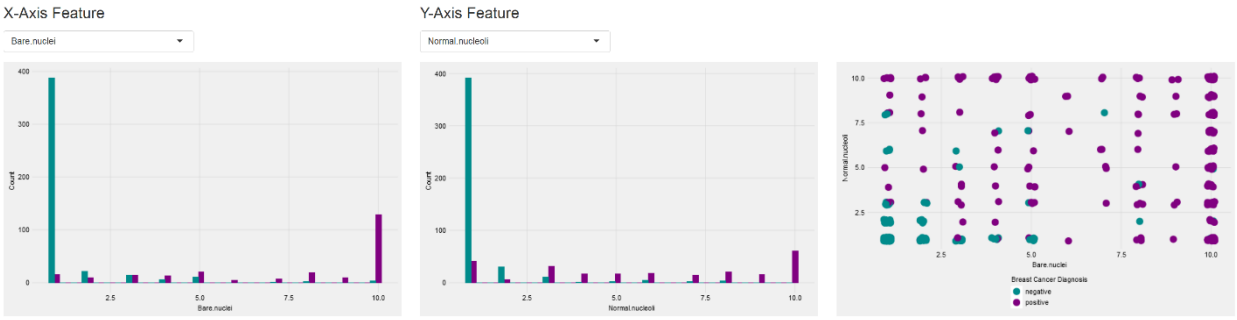

Figure10 Relation Between "Bare.nuclei" and "Normal.nucleoli"

Normal Nucleoli are small structures that can be seen in cells' nucleus. As shown in figure10, the value of Normal Nucleoli in normal cells is small because, usually, the cell nucleolus is very small if visible comparing to Malignant cells.

Experiment2. At the first time the application run as shown in figure11, the k value (NO. of neighbors) is tuned to be 5 as a default value

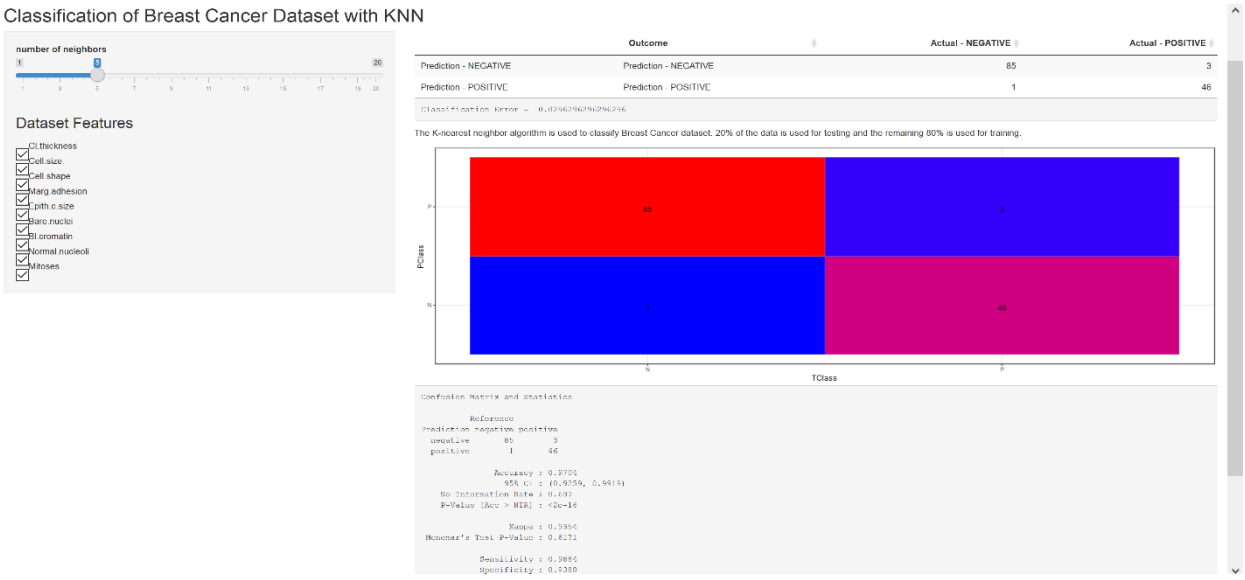

Figure11 Classification Results With k=5 (Default Value)

It is worthy to mention that as this tool will be used in breast cancer detection, the goal here is to select the value of k and features that achieve the minimum false negative FNs as mentioned earlier. When the value of k is equal to 5, and all features are selected, NO. of FN is 3 with accuracy equal to 97% and a classification error

3%. Now the value of k will be changed to check whether 5 is the optimal number of the nearest neighbors or not.
When the value of k is tuned to be 2,7,10,11 the number of FNs are 3,4,2 and 2 with accuracy equal to 95.5%, 95.5%, 97% and 97.78% respectively. Classification error values are 4%, 4%, 3% and 2% respectively. From these results, it can be concluded that the performance of the classifier is the optimum with k value equal to 11.
Experiment3. After choosing the optimal k value that achieves the best classification accuracy and minimum FN number, in this experiment, each attribute will be evaluated with respect to the prediction class to inspect how much information that an attribute can give about the class (information gain). This can be accomplished by removing one attribute at a time and run the classifier to evaluate its performance with the absence of that removed attribute. Before starting this experiment, the information gain for each attribute is measured and ranked by their individual evaluation as presented in table3.

Table3 Ranked Information Gain For all Features

Rank	Feature	Feature Importance
1	"Cell.size"	0.46
2	"Cell.shape"	0.45
3	"Bare.nuclei"	0.40
4	"Bl. Chromatin"	0.37
5	"Epithelial. c. size"	0.35
6	"Normal.nucleoli"	0.32
7	"Cl. thickness"	0.30
8	"Marg.adhesion"	0.30
9	"Mitoses"	0.14

Table4 presents the results of this experiment. Note: since that k=11 achieved the minimum error and FN rate, the value of k in this experiment is chosen to be 11.

NO. of attributes	Attribute removed	accuracy	Classification Error	NO. of FN
All 9 attributes	-	97.78%	2%	2
8 attributes	"Cell.size"	96%	4%	4
8 attributes	"Cell.shape"	96%	4%	4
8 attributes	"Bare.nuclei"	95.8%	5%	5
8 attributes	"Bl. Chromatin"	96%	4%	4
8 attributes	"Epithelial. c. size"	97.78%	2%	2
8 attributes	"Normal.nucleoli"	95.56%	4%	4
8 attributes	"Cl. thickness"	96%	4%	4
8 attributes	"Marg.adhesion"	96%	4%	3
8 attributes	"Mitoses"	97%	3%	3

Table 4 Comparison of Classifier Performance after Removing One Attribute at a Time

From the previous table and from comparing the error rate after removing each feature, it can be noticed that the attribute "Epithelial Cell Size" has the least effect on the classifier performance with the same accuracy, classification error and number of FNs of all-attributes classification. The second attribute that has less impact on the classifier performance is "Mitosis" with just .78 difference from classification with the whole nine attributes. Removing the attributes "Normal Nucleoli" and "Bare Nucleoli" generated the lowest accuracy with 95.56% and 95.8% and the highest error rate with 4% and 5% respectively.
Experiment4. In this experiment, at the first time the app run, these parameters are tuned to be 10 for C and 1 for Gamma. With these values, the SVM classifier achieved 91.42% accuracy and 88.76% sensitivity as illustrated in fugure12.

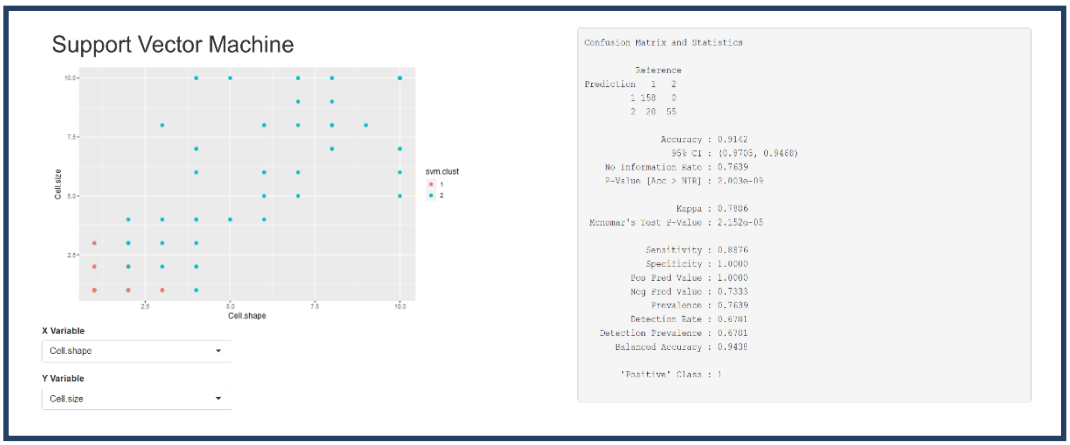

Figure12 Classification Results With C=10, Gamma=1 (Default Value)

After increasing Gamma parameter to 2.1, it is noticed that the accuracy and sensitivity decreased to 83.26%and 78% respectively. If the value continued to be decreased gradually, the accuracy will be decreased as well, thus, it can be concluded that reducing Gamma will negatively affect the classification accuracy. Increasing and decreasing the value of C parameter will change the margin around the separation line, i.e. the SVM optimizer, based on this value, will consider smaller or larger margin around the separation line even if that line will misclassify more data points. The best performance achieved is with accuracy equal to 97.85% and 97.19% sensitivity when C equal to 8.6 and Gamma equal to .15.

Experiment5. In this experiment, the Naïve Bayes algorithm is applied on the dataset in order to examine its classification performance. As shown in figure13, the classifier achieved the highest accuracy with 97.78%. and sensitivity 97.70%

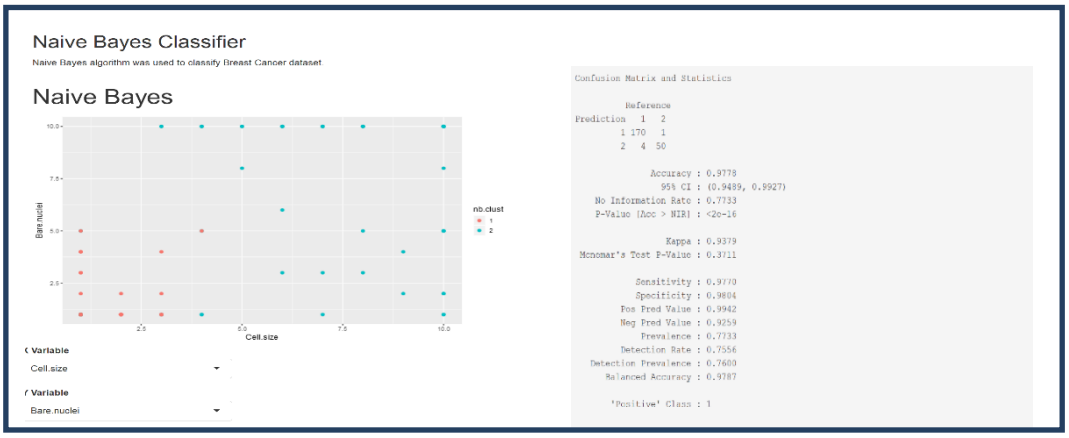

Figure13 NB Classification Results

5- Future Work

With user-friendly interfaces and by allowing the user to enter features values, this interactive data analysis and visualization tool and similar real-time data visualization tools can be used by doctors and scientists in medical labs and even non-technical users to precisely detect and diagnosis fatal diseases, in which early detection is critically significant such as cancers, by providing an enhanced computational analysis and interpretation thus saving patients' lives.

6 – Conclusion

This study presents an interactive data analysis and visualization tool that allow the user to interactively compare the performance of three different ML algorithms (KNN, SVM and NB) in terms of accuracy, sensitivity and error rate on Wisconsin Breast Cancer datasets. The user also can visually adjust the

classifiers' parameters such as the value of K in KNN and Cost, Gamma in SVM to gain the maximum accuracy in a user-friendly way. The experiment results indicate that all the selected algorithms significantly presented high performance on determining whether the breast lump is benign or malignant. Although, the results are approximated, SVM has proven its efficiency and achieves the best performance with 97.85% accuracy.

References

[1] D. Lobach *et al.*, "Evidence Report/Technology Assessment Enabling Health Care Decisionmaking Through Clinical Decision Support and Knowledge Management," 2012.
[2] I. Ko and H. Chang, "Interactive data visualization based on conventional statistical findings for antihypertensive prescriptions using National Health Insurance claims data," *Int. J. Med. Inform.*, vol. 116, no. February, pp. 1–8, 2018.
[3] C. Plaisant, M. Monroe, T. Meyer, and B. Shneiderman, "Interactive Visualization," *Big Data Heal. Anal.*, pp. 1–18, 2014.
[4] B. Shneiderman, C. Plaisant, and B. W. Hesse, "Improving healthcare with interactive visualization," *Computer (Long. Beach. Calif).*, vol. 46, no. 5, pp. 58–66, 2013.
[5] "Figure 1. Proposed Breast Cancer Diagnosis Model TABLE 1 DESCRIPTION OF THE BREAST CANCER DATASETS," 2011.
[6] "Wisconsin Breast Cancer (Diagnostic) DataSet Analysis." [Online]. Available: http://rstudio-pubs-static.s3.amazonaws.com/344010_1f4d6691092d4544bfbddb092e7223d2.html. [Accessed: 05-Nov-2018].
[7] "Breast cancer statistics | World Cancer Research Fund." [Online]. Available: https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics. [Accessed: 18-Nov-2018].
[8] P. Mendoza, M. Lacambra, P.-H. Tan, and G. M. Tse, "Fine needle aspiration cytology of the breast: the nonmalignant categories.," *Patholog. Res. Int.*, vol. 2011, p. 547580, May 2011.
[9] D. A. Ellis and H. L. Merdian, "Thinking outside the box: Developing dynamic data visualizations for psychology with Shiny," *Front. Psychol.*, vol. 6, no. DEC, pp. 1–6, 2015.
[10] R. Agrawal, A. Kadadi, X. Dai, and F. Andres, "Challenges and opportunities with big data visualization," *Proc. 7th Int. Conf. Manag. Comput. Collect. Intell. Digit. Ecosyst. - MEDES '15*, no. October, pp. 169–173, 2015.
[11] W. Cho, Y. Lim, H. Lee, M. K. Varma, M. Lee, and E. Choi, "Big Data Analysis with Interactive Visualization using R packages," *Proc. 2014 Int. Conf. Big Data Sci. Comput. - BigDataScience '14*, pp. 1–6, 2014.
[12] L. Wang, G. Wang, and C. A. Alexander, "Big Data and Visualization: Methods, Challenges and Technology Progress," *Digit. Technol.*, vol. 1, no. 1, pp. 33–38, 2015.
[13] "Markus Loecher and Karl Ropkins (2015). RgoogleMaps and loa: Unleashing R Graphics Power on Map Tiles. Journal of Statistical Software 63(4), 1-18."
[14] M. Gesmann and D. De Castillo, "Using the Google Visualisation API with R."
[15] "Introducing Shiny: Easy web applications in R | RStudio Blog." [Online]. Available: https://blog.rstudio.com/2012/11/08/introducing-shiny/. [Accessed: 18-Oct-2018].
[16] O. Scrivner, V. Chakilam, J. Poojary, N. Sahoo, C. Uppuluri, and S. De Spiegeleire, "Building Customized Text Mining Tools via Shiny Framework: The Future of Data Visualization," no. May 2017.

[17] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, 2016.
[18] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability," *Int. J. Comput. Sci. Mob. Comput.*, vol. 3, no. 1, pp. 10–22, 2014.
[19] S. Aruna, S. P. Rajagopalan, L. V Nandakishore, and S. C. In, "Knowledge Based Analysis Of Various Statistical Tools In Detecting Breast Cancer," pp. 37–45, 2011.
[20] L. Rodrigues, "Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection," *XI Work. Visão Comput.*, no. December, pp. 415–423, 2016.
[21] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original). [Accessed: 18-Nov-2018].
[22] E. Beauxis-aussalet and L. Hardman, "Simplifying the Visualization of Confusion Matrix," no. May, pp. 1–2, 2016.
[23] "Understanding Confusion Matrix – Towards Data Science." [Online]. Available: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62. [Accessed: 19-Nov-2018].